Optimal Bernoulli Routing in an Unreliable $M/G/1$ Retrial Queue

Nathan P. Sherman
Directorate of Force Management Policy
U.S. Air Force Headquarters, Manpower and Personnel
1040 Air Force Pentagon
Washington, DC 20330-1040, USA
Email: nathan.sherman@us.af.mil

and

Jeffrey P. Kharoufeh[1]
Department of Industrial Engineering
University of Pittsburgh
1048 Benedum Hall
3700 O'Hara Street
Pittsburgh, PA 15261, USA
Email: jkharouf@pitt.edu

Final version to appear in
Probability in the Engineering and Informational Sciences.

Abstract

Recently, Sherman et al. [14] analyzed an $M/G/1$ retrial queueing model in which customers are forced to retry their service if interrupted by a server failure. Using classical techniques, they provided a stability analysis, queue length distributions, key performance parameters, and stochastic decomposition results. We analyze the system under a static Bernoulli routing policy that routes a proportion of arriving customers directly to the orbit when the server is busy or failed. In addition to providing the key performance parameters, we show that this system exhibits a dual stability structure, and we characterize the optimal Bernoulli routing policy that minimizes the total expected holding costs per unit time.

Keywords: Retrial queue, unreliable server, Bernoulli routing.

[1]Author to whom correspondence should be sent.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Optimal Bernoulli Routing in an Unreliable M/G/1 Retrial Queue	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Air Force Headquarters, Manpower and Personnel,Directorate of Force Management Policy,1040 Air Force Pentagon,Washington,DC,20330-1040**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Probability in the Engineering and Informational Sciences, 25 (1), 1-20

14. ABSTRACT

Recently, Sherman et al. [14] analyzed an M=G=1 retrial queueing model in which customers are forced to retry their service if interrupted by a server failure. Using classical techniques they provided a stability analysis, queue length distributions, key performance parameters, and stochastic decomposition results. We analyze the system under a static Bernoulli routing policy that routes a proportion of arriving customers directly to the orbit when the server is busy or failed. In addition to providing the key performance parameters, we show that this system exhibits a dual stability structure, and we characterize the optimal Bernoulli routing policy that minimizes the total expected holding costs per unit time.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**19**	

1 Introduction

As a model for streaming multimedia applications, Sherman, et al. [14] recently analyzed an $M/G/1$ retrial queueing system with an unreliable server, infinite-capacity retrial queue (or orbit), and an infinite-capacity primary queue. In that model, a customer in service is forced to join the orbit if the server fails during his/her service cycle. Customers sent to the orbit persistently and independently retry the server at random intervals until their service is complete, but they can only regain access to the server if it is up and idle at the time of a retrial attempt. Primary customer arrivals who find the server busy or failed always join the FIFO primary queue, whereas those who find the server up and idle are served to completion if their service is not interrupted by a server failure. Using the method of supplementary variables, they established the existence of dual stability conditions, one for the orbit and one for the primary queue, and derived the probability generating functions of the primary queue length and orbit size, as well as the mean performance parameters. They also demonstrated that the orbit size and system size exhibit a stochastic decomposition property. Their model assumes that arriving customers join the primary queue by default if the server is found busy or failed. However, in many engineering applications, it is possible to mitigate congestion in the primary queue by statically or dynamically routing some arriving customers directly to the orbit to retry their service later. The primary aim of this extended note is to analyze the unreliable retrial model of [14] when a controller uses a Bernoulli routing policy to manage congestion in the primary queue.

Our model here is motivated by a particular type of internet protocol for streaming media applications known as *IP multicast*, or simply *multicast*. Multicast provides a way to deliver a single media stream to a group of users linked via a local area network (LAN). For example, a training seminar being conducted at a particular location can be streamed to groups of individuals on a LAN at a separate, remote location. A digitized camera feed uploads the live video to the internet, and recipients participate in the seminar (though not interactively) through the streamed video at individual computer network terminals. To ensure wider dissemination of the seminar, it is often desirable to save a complete copy of the streamed content on a hard disk at the remote location for future playback. Because the seminar is live, there is a need to ensure timely transmission of real-time packets, even at the expense of some packet losses. Packets may be dropped by the network administrator to relieve congestion in the primary packet transmission queue, or they can be dropped if their transmission fails due to packet corruption, hardware failures, software errors, or congestion in the local network or the internet itself. However, dropped packets can be retransmitted later (when the transmission medium becomes free) so that the complete seminar can be "patched-up" in the stored copy for future playback. Consequently, dropped packets are not lost but are necessary to ensure a high-quality stored copy. The primary packet transmission queue mimics a 1-persistent carrier-sense multiple-access (CSMA) system. When the oldest packet in the primary queue detects that the transmission medium is free, transmission begins immediately. If the communication medium fails during transmission, the packet is sent to a retrial queue which is

analogous to a non-persistent CSMA system. In a non-persistent system, packets do not persist to wait for a free transmission medium but rather retry the transmission medium at random intervals until it is found up and idle. Therefore, those packets that are dropped, either upon arrival by the network administrator or due to a server failure, are assigned a lower priority than the real-time packets that enter the primary queue and are time-sensitive. The latter type corresponds to the primary (or priority) customers. Because dropped or interrupted packets can be retransmitted for inclusion in the stored copy, their transmission time is no longer important. These packets correspond to the retrial customers who enter the orbit and retry the server until it is found up and idle. The administrator seeks to determine the optimal (static) proportion of arriving packets to admit to the primary queue with the objective of minimizing the total expected holding costs per unit time in the primary queue and the orbit.

While it is well known that dynamic routing policies are well-suited for non-stationary regimes and generally outperform static policies, dynamic routing requires a great deal of information gathering, storage, updating, and exchange. By contrast, static routing can be much faster, easier to implement, and require far less overhead. Applications in computer/communications networks and service systems abound. For example, Combeé and Boxma [7] considered the allocation of tasks to a finite number of distributed processors. In ATM networks, static routing algorithms are often employed to balance the load across computer network links (cf. Casseti et al. [4]). In service systems, Servi and Humair [12] optimized Bernoulli routing probabilities at discrete time points to balance the load in large-scale call centers by estimating arrival rates. Besides their relatively low overhead requirements, static policies can also be used to aid engineers in designing communications or computer networks by providing performance bounds that can be easily computed and evaluated.

More generally, Bernoulli routing of arriving customers to a finite group of ordinary (non-retrial), homogeneous queueing systems has been studied extensively in the queueing literature (cf. [5, 10] and references therein). Under certain conditions, it can be shown that the optimal Bernoulli routing policy is to assign an arriving customer to any one of a finite number of available servers with equal probability. Relatively few researchers have considered routing policies for arriving customers in retrial systems. Choi and Park [6] analyzed a Bernoulli routing policy for a system that is similar to the model in [14] except that it does not consider server failures. Atencia and Moreno [3] examined Bernoulli routing in a model with general (as opposed to exponential) retrial times. Their model permits only the retrial customer at the head of the line to retry the server which is assumed to be reliable. Liang and Kulkarni [11] studied optimal dynamic routing in a retrial system in which both retrial customers and primary arrivals are routed either to the primary queue or to the orbit. They proved the existence of a threshold-type policy that routes all customers to the primary queue up to a threshold after which all arrivals are routed to the orbit. Their model, however, does not consider unreliability of the server which significantly complicates the analysis.

In this extended note, we revisit the queueing system analyzed in [14] and include static Bernoulli routing of arriving customers who find the server busy or failed. Specifically, a con-

troller routes an arriving customer to the primary queue with (fixed) probability q and to the orbit with complementary probability $1-q$, independently of everything else, during busy or failed periods. Our main objective is to provide the primary performance parameters of this system and to determine the optimal Bernoulli routing policy that minimizes the total expected holding costs per unit time. Following an analysis similar to the one in [14], we (i) provide the necessary and sufficient overall stability condition, (ii) derive the generating functions of the primary queue length and the orbit size distributions, (iii) provide the mean congestion performance parameters, and (iv) determine the optimal Bernoulli routing policy. It is shown that the stability condition of the overall system and the steady state distribution of the server's status are insensitive to the Bernoulli routing parameter; however, the routing parameter plays a crucial role in establishing the stability region of the primary queue. Moreover, we provide sufficient conditions to ensure the existence of a unique, optimal Bernoulli routing policy that minimizes the expected total cost per unit time of holding customers in the primary queue and the orbit. The structure of the cost function is characterized explicitly by the holding cost coefficients, the service time distribution and the arrival, failure, retrial, and repair rates.

The remainder of the paper is organized as follows. Section 2 provides a complete model description and defines the key generating functions. In section 3, we state the necessary and sufficient condition for overall system stability and provide the generating functions of queue lengths when the server is idle, busy, or failed. We also provide expressions for the mean queue lengths needed for the cost function. Section 4 shows how our model generalizes a few other $M/G/1$ retrial models in the literature. Section 5 provides structural results for the cost function and characterizes the optimal Bernoulli routing policy, while section 6 provides a few illustrative examples.

2 Model Description

Primary customers arrive to the queueing system in accordance with a homogeneous Poisson process with rate λ. A primary customer who finds the server idle (and not under repair) seizes the server and is completed if no server failures occur during its service cycle. Should the server fail during service, the interrupted customer joins a retrial queue (or orbit) from which it persistently retries the server at random intervals until access is regained. Customers who are interrupted must repeat their service cycle (i.e., a preemptive repeat discipline is employed). In contrast to the model of [14], the system uses a controller who diverts a proportion of arriving primary customers directly to the retrial orbit when the server is busy or failed to manage the congestion level of the primary queue. If a primary arrival finds the server busy or under repair, the controller routes the customer to the primary queue with probability q ($0 \leq q \leq 1$), and to the retrial orbit with complementary probability $p = 1-q$, independently of everything else (i.e., a Bernoulli routing policy is used). The uninterrupted service times, $\{S_n : n \geq 1\}$, are independent and identically distributed (i.i.d.) with absolutely continuous cumulative distribution function (c.d.f.) B and probability density function

(p.d.f.) b. For $s \geq 0$, let

$$b^*(s) = \int_0^\infty e^{-sx} b(x) \mathrm{d}x$$

denote the Laplace transform of b. We consider here both active and idle failures of the server. That is, server failures occur according to a Poisson process with rate ξ whenever the server is idle or busy; however, the server cannot fail when it is under repair. The repair time is assumed to be exponentially distributed with mean $1/\alpha$. We pause here to note that, while it may be possible to analyze the model with general repair times (by adding a supplementary variable or by embedding a Markov chain at appropriate epochs), we assume exponential repair times to maintain transparency of the analysis and consistency with the model analyzed in [14]. Customers who enter the retrial orbit (either by virtue of a server failure or by being routed to the orbit by the controller) retry the server directly at exponentially-distributed time intervals with mean $1/\theta$. The inter-retrial times are independent and identically distributed for each customer in the retrial group. Moreover, retrial customers behave independently of one another, of customers in the primary queue, and of external arrivals to the system. Finally, customers in the retrial group can gain access to the server only if it is up and idle at the time of a retrial attempt. The arrival, service, failure, repair, and retrial processes are assumed to be mutually independent.

As in Sherman et al. [14], we analyze this model using classical techniques, namely the method of supplementary variables and probability generating functions. Adopting their notation, let Q_t denote the number of customers in the primary queue at time t, excluding any customer that might be in service, and let R_t denote the number of customers in the retrial group at time t. The random variable U_t is the occupation status of the server given by

$$U_t = \begin{cases} 1, & \text{if the server is occupied at time } t, \\ 0, & \text{if the server is not occupied at time } t, \end{cases}$$

while S_t describes the operational status of the server at time t defined by

$$S_t = \begin{cases} 1, & \text{if the server is not failed at time } t, \\ 0, & \text{if the server is failed at time } t. \end{cases}$$

Let X_t denote the elapsed service time of the customer in service at time t so that the continuous-time stochastic process, $\{(Q_t, U_t, R_t, S_t, X_t) : t \geq 0\}$ is a Markov process describing the state of the system. Further define N_t as the total number of customers in the system at time t (i.e., in orbit, in the primary queue, and in service). Our primary aim is to study the steady state versions of Q_t, R_t and N_t which we denote by Q, R, and N, respectively. Using these quantities, we will establish conditions under which a unique optimal Bernoulli routing parameter exists. However, before doing so, it is of interest to examine the influence of Bernoulli routing on the stability condition, queue length distributions, and performance parameters of the queueing system. To this end, define for

$j \geq 0$, $k \geq 0$, and $x \geq 0$,

$$\begin{aligned}
\pi_{0,0,j,1} &= \lim_{t\to\infty} \mathbb{P}(Q_t = 0, U_t = 0, R_t = j, S_t = 1), \\
\pi_{k,0,j,0} &= \lim_{t\to\infty} \mathbb{P}(Q_t = k, U_t = 0, R_t = j, S_t = 0), \\
\pi_{k,1,j,1}(x) &= \lim_{t\to\infty} \mathbb{P}(Q_t = k, U_t = 1, R_t = j, S_t = 1, X_t \leq x),
\end{aligned}$$

the limiting probabilities that the server is idle, failed, or busy, respectively, when there are j customers in the retrial group and k customers in the primary queue. Next, define the generating functions

$$\phi_{0,0,1}(z_1) = \sum_{j=0}^{\infty} z_1^j \pi_{0,0,j,1}, \quad |z_1| \leq 1,$$

$$\phi_{k,0,0}(z_1) = \sum_{j=0}^{\infty} z_1^j \pi_{k,0,j,0}, \quad |z_1| \leq 1,$$

$$\phi_{k,1,1}(x, z_1) = \sum_{j=0}^{\infty} z_1^j \pi_{k,1,j,1}(x), \quad |z_1| \leq 1,\ x \geq 0.$$

Here, $\phi_{0,0,1}(z_1)$ is the generating function of R when the server is idle, $\phi_{k,0,0}(z_1)$ is the generating function of R when the server is failed and k customers are awaiting service in the primary queue, and $\phi_{k,1,1}(x, z_1)$ is the generating function of R when the server is busy, k customers are in the primary queue, and the elapsed service time of the customer in service has not exceeded x. Further define, respectively,

$$\psi_{0,0}(z_1, z_2) = \sum_{k=0}^{\infty} z_2^k \phi_{k,0,0}(z_1), \quad |z_1| \leq 1,\ |z_2| \leq 1,$$

$$\psi_{1,1}(x, z_1, z_2) = \sum_{k=0}^{\infty} z_2^k \phi_{k,1,1}(x, z_1), \quad |z_1| \leq 1,\ |z_2| \leq 1,$$

the generating functions of $\phi_{k,0,0}(z_1)$ and $\phi_{k,1,1}(x, z_1)$ with respect to the primary queue size. The joint generating function of the orbit and primary queue size when the server is busy is given by

$$\psi_{1,1}(z_1, z_2) = \int_0^{\infty} \psi_{1,1}(x, z_1, z_2) \mathrm{d}x.$$

The joint generating function of (R, Q) will be denoted by

$$G(z_1, z_2) \equiv \mathbb{E}\left(z_1^R, z_2^Q\right) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \mathbb{P}(R = j, Q = k)\, z_1^j\, z_2^k, \quad |z_1| \leq 1,\ |z_2| \leq 1,$$

and the generating function of the steady state system size N is given by

$$H(z) \equiv \mathbb{E}\left(z^N\right) = \sum_{j=0}^{\infty} \mathbb{P}(N = j)\, z^j, \quad |z| \leq 1.$$

In section 3, we provide the generating functions $\phi_{0,0,1}(z_1)$, $\psi_{0,0}(z_1, z_2)$, and $\psi_{1,1}(z_1, z_2)$ when a Bernoulli routing policy is used. The relevant performance parameters are derived from these results.

3 Main Results

In this section, we state the necessary and sufficient condition for stability of the overall queueing system and also provide the generating functions of (R, Q) and N. Central to the analysis is the notion of the fundamental server period introduced by Aissani and Artalejo [1]. The fundamental server period is the time from which the server initiates a new service cycle until the next time it commences a new service cycle. Let N_r and N_q respectively denote the number of customers entering the orbit and primary queue during a fundamental server period, and let $a(i,j) = \mathbb{P}(N_r = i, N_q = j)$, $i, j \geq 0$. As before, we let

$$Q(z_1, z_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a(i,j) z_1^i z_2^j, \qquad |z_1| \leq 1, |z_2| \leq 1.$$

Following along the lines of the proof of Theorem 2.1 of [1], it can be verified that the joint generating function of (N_r, N_q) is given by

$$Q(z_1, z_2) = \hat{B}(z_1, z_2) + \frac{\alpha z_1 \xi (1 - \hat{B}(z_1, z_2))}{(\alpha + \lambda(1 - pz_1 - qz_2))(\xi + \lambda(1 - pz_1 - qz_2))}$$

where $\hat{B}(z_1, z_2) \equiv b^*(\xi + \lambda(1 - pz_1 - qz_2))$ and $b^*(x)$ is the Laplace transform of b evaluated at x. The first term on the right-hand side of $Q(z_1, z_2)$ is the generating function of the total number of arrivals to the system when there is no server failure during the service cycle that begins the fundamental server period, while the second term is the generating function of the same quantity when the service is interrupted by a failure and the server undergoes a repair. When $q = 1$ $(p = 0)$, $Q(z_1, z_2)$ reduces to the generating function of (N_r, N_q) of [14]. Using standard methods, the expected number of arrivals to the primary queue during a fundamental server period is given by

$$\rho_1 = Q'_{z_2}(1,1) = \frac{\lambda q (1 - b^*(\xi))(\alpha + \xi)}{\alpha \xi},$$

where $Q'_{z_2}(z_1, z_2) \equiv \partial Q(z_1, z_2)/\partial z_2$. In what follows, it will be shown that ρ_1 plays an important role in determining the stability condition of the primary queue. Moreover, it can be shown (see Lemma 1 in [14]) that the equation

$$z_2 - Q(z_1, z_2) = 0$$

has a unique solution, call it $g(z_1)$, inside the region $|z_2| < 1$ whenever $|z_1| < 1$ or $|z_1| \leq 1$ and $\rho_1 > 1$. Whenever $z_1 = 1$, $g(1)$ is the smallest positive real zero with $g(1) < 1$ if $\rho_1 > 1$, and $g(1) = 1$ if $\rho_1 \leq 1$.

Next, we state the necessary and sufficient condition for stability of the overall system (and the orbit) and simultaneously state the generating functions $\phi_{0,0,1}(z_1)$, $\psi_{0,0}(z_1, z_2)$, and $\psi_{1,1}(z_1, z_2)$. These results generalize those reported in [9, 14].

Theorem 1 *The queueing system is stable if and only if*

$$\rho \equiv \frac{\lambda(1 - b^*(\xi))(\alpha + \xi)}{\alpha b^*(\xi) \xi} < 1. \tag{1}$$

When $\rho < 1$, the generating functions $\phi_{0,0,1}(z_1)$, $\psi_{0,0}(z_1, z_2)$, and $\psi_{1,1}(z_1, z_2)$ are given by

$$\phi_{0,0,1}(z_1) = (1-\rho)\left(\frac{\alpha}{\alpha+\xi}\right)\exp\left\{-\frac{1}{\theta}\int_{z_1}^{1}\frac{\lambda(1-g(u))+\xi\left(1-\frac{\alpha}{\alpha+\lambda(1-pu-qg(u))}\right)}{g(u)-u}\,du\right\}, \quad (2)$$

$$\psi_{0,0}(z_1, z_2) = \xi\,\frac{\Lambda_{0,0}(z_1, z_2)}{\Lambda(z_1, z_2)}\,\phi_{0,0,1}(z_1), \quad (3)$$

and

$$\psi_{1,1}(z_1, z_2) = \lambda\,\frac{\Lambda_{1,1}(z_1, z_2)}{\Lambda(z_1, z_2)}\,\phi_{0,0,1}(z_1) \quad (4)$$

where

$$\begin{aligned}\Lambda_{0,0}(z_1, z_2) = {} & (g(z_1)-z_1)[\alpha+\lambda(1-pz_1-qg(z_1))][\xi+\lambda(1-pz_1-qz_2)][z_2-\hat{B}(z_1, z_2)-z_1(1-\hat{B}(z_1, z_2))] \\ & + \lambda z_1(1-\hat{B}(z_1, z_2))(z_2-z_1)(1-pz_1-qg(z_1))[\alpha+\xi+\lambda(1-pz_1-qg(z_1))],\end{aligned}$$

$$\begin{aligned}\Lambda_{1,1}(z_1, z_2) = {} & (1-\hat{B}(z_1, z_2))(z_2-g(z_1)) \\ & \times[(1-z_1)\{\alpha\xi+[\alpha+\lambda(1-pz_1-qz_2)][\alpha+\xi+\lambda(1-pz_1-qg(z_1))]\} \\ & +\lambda\xi(1-pz_1-qg(z_1))(1-pz_1-qz_2)],\end{aligned}$$

$$\begin{aligned}\Lambda(z_1, z_2) = {} & (g(z_1)-z_1)[\alpha+\lambda(1-pz_1-qg(z_1))] \\ & \times\left[(z_2-\hat{B}(z_1, z_2))\,[\alpha+\lambda(1-pz_1-qz_2)]\,[\xi+\lambda(1-pz_1-qz_2)]-\alpha\xi(1-\hat{B}(z_1, z_2))z_1\right],\end{aligned}$$

and for $z_1 \in [0, 1]$, $g(z_1)$ verifies

$$g(z_1) = b^*(\xi+\lambda(1-pz_1-qg(z_1))) + \frac{\alpha\xi z_1[1-b^*(\xi+\lambda(1-pz_1-qg(z_1)))]}{[\alpha+\lambda(1-pz_1-qg(z_1))][\xi+\lambda(1-pz_1-qg(z_1))]}.$$

Proof. We omit the proof for the sake of brevity. However, the result can be shown by following the steps of the proof of Theorem 1 in [14]. Note that the generating function $\hat{B}(z_2)$ in [14] is replaced by $\hat{B}(z_1, z_2)$ above. ∎

If a Bernoulli routing scheme is not used (i.e., if $q = 1$), all arriving customers who find the server busy or under repair will join the primary queue by default. In such case, the generating functions (2)–(4) are the same as those reported in Theorem 1 of [14].

The system controller is faced with the task of deciding the appropriate proportion of arriving customers to route to the retrial orbit when the server is busy or failed. If this proportion is too small, the primary queue may become unstable. On the other hand, if the proportion is too large, a significant number of customers will be denied immediate access to the service system and asked to return later for service. Our objective is to determine the Bernoulli routing policy that balances this tradeoff and minimizes the total expected holding costs per unit time. To this end, we next characterize the primary performance parameters (namely the mean queue lengths) using the generating functions of (R, Q) and N.

Proposition 1 *For $\rho < 1$, the generating function of (R, Q) is given by*

$$G(z_1, z_2) = \phi_{0,0,1}(z_1) \left[1 + \xi \frac{\Lambda_{0,0}(z_1, z_2)}{\Lambda(z_1, z_2)} + \lambda \frac{\Lambda_{1,1}(z_1, z_2)}{\Lambda(z_1, z_2)} \right], \tag{5}$$

and the generating function of N is given by

$$H(z) = \phi_{0,0,1}(z) \, \frac{\Psi_1(z)}{\Psi_2(z)}, \tag{6}$$

where

$$\Psi_1(z) = [\alpha + \xi + \lambda - \lambda z] b^*(\xi + \lambda - \lambda z) \left\{ \alpha\xi + \lambda(1 - z)[\alpha + \xi + \lambda - \lambda z] \right\},$$

$$\Psi_2(z) = [\alpha + \lambda - \lambda z] \left\{ \alpha\xi b^*(\xi + \lambda - \lambda z) - \lambda(z - b^*(\xi + \lambda - \lambda z))[\alpha + \xi + \lambda - \lambda z] \right\},$$

and $\Lambda_{0,0}(z_1, z_2)$, $\Lambda_{1,1}(z_1, z_2)$, and $\Lambda(z_1, z_2)$ are defined in Theorem 1.

Proof. The generating function of (R, Q) is obtained by summing over the three mutually exclusive and exhaustive server states, i.e.,

$$G(z_1, z_2) = \phi_{0,0,1}(z_1) + \psi_{0,0}(z_1, z_2) + \psi_{1,1}(z_1, z_2)$$

where $\phi_{0,0,1}(z_1)$, $\psi_{0,0}(z_1, z_2)$, and $\psi_{1,1}(z_1, z_2)$ are given by (2), (3), and (4), respectively. Similarly, the generating function $H(z)$ is obtained directly by

$$H(z) = \phi_{0,0,1}(z) + \psi_{0,0}(z, z) + z \, \psi_{1,1}(z, z).$$

■

Remark: When $q = 1$, these results are the same as those reported in Corollary 1 of [14]. Alternatively, if we assume that the server is reliable (i.e., if we allow $\xi \downarrow 0$), then we obtain the joint generating function of Choi and Park [6]. Next, we obtain the mean values of R, Q, and N in the following proposition.

Proposition 2 *The steady state mean orbit size, mean primary queue size, and mean number in system are respectively given by*

$$\mathbb{E}(R) = \frac{1}{1 - \rho} \left[\frac{\lambda\xi p + \alpha\rho(\xi + \lambda p)}{\alpha\theta} \right.$$
$$\left. + \frac{\lambda}{\xi \, b^*(\xi)} \cdot \frac{(1 - \rho_1/\rho) \left\{ \xi^3 b^*(\xi) - (\alpha + \xi)^2 [\xi \hat{B}' - \lambda(1 - b^*(\xi))] \right\} + \alpha\xi^2 b^*(\xi)(1 - b^*(\xi))(1 - \rho_1)}{\alpha\xi b^*(\xi)(\alpha + \xi)(1 - \rho_1)} \right], \tag{7}$$

$$\mathbb{E}(Q) = \frac{\lambda q}{1 - \rho_1} \left[\frac{\xi^2 b^*(\xi) - (\alpha + \xi)[(\alpha + \xi)\hat{B}' - \alpha\rho b^*(\xi)]}{\alpha\xi b^*(\xi)(\alpha + \xi)} \right] \tag{8}$$

and

$$\mathbb{E}(N) = \frac{1}{1-\rho}\left[\frac{\lambda\xi p + \alpha\rho(\xi+\lambda p)}{\alpha\theta} + \frac{\lambda}{\xi\, b^*(\xi)} \cdot \frac{b^*(\xi)\left\{\xi^3 + (1-b^*(\xi))[\alpha\xi(\alpha+2\xi) + \lambda(\alpha+\xi)^2]\right\} - \xi(\alpha+\xi)^2\hat{B}'}{\alpha\xi b^*(\xi)(\alpha+\xi)}\right] \quad (9)$$

where $\hat{B}' = \lambda\int_0^\infty xe^{-\xi x}b(x)dx$.

Equations (7) and (9) show that $\rho < 1$ is necessary to ensure the stability of the retrial orbit and the system, and by (8) we see that $\rho_1 < 1$ is necessary for the stability of the primary queue. Clearly, for any $\xi \geq 0$, $\rho_1 \leq \rho$ so that the stability of R depends on ρ and not ρ_1. Moreover, when $\xi > 0$ and $q < 1$, some interesting insights about the stability of the primary queue and the orbit are revealed. Specifically, there exists a set of model parameters for which the primary queue will remain stable even if the orbit is not stable. The dynamics of the system dictate that retrial customers become subordinate to primary customers since they can regain access to the server only when it is found to be up and idle. That is, retrial customers experience a smaller effective service rate than do primary customers; therefore, the orbit may continue to grow while the primary queue remains stable. From a service quality perspective, it is desirable to keep both average queue lengths small; however, under certain conditions, the controller may choose to route arrivals only to the primary queue, or only to the orbit, during busy or down periods. Next, using the results of Theorem 1, the steady state distribution of the server's status is obtained.

Proposition 3 *For $\rho < 1$, the steady state distribution of the server's status is given by*

$$\mathbb{P}(\text{Idle}) = \phi_{0,0,1}(1) = (1-\rho)\left(\frac{\alpha}{\alpha+\xi}\right),$$

$$\mathbb{P}(\text{Failed}) = \psi_{0,0}(1,1) = \frac{\xi}{\alpha+\xi},$$

and

$$\mathbb{P}(\text{Busy}) = \psi_{1,1}(1,1) = \rho\left(\frac{\alpha}{\alpha+\xi}\right).$$

Remark: The steady state distribution of the server's status is intuitive, i.e., the probability that the server is busy is simply the traffic intensity ρ scaled by the long-run proportion of time that the server is not under repair, $\alpha(\alpha+\xi)^{-1}$. Similarly, the steady state probability that the server is idle is $(1-\rho)$ scaled by the proportion of time the server is not under repair. Because our model allows for both active and idle server failures, the long-run proportion of time the server is failed is intuitively given by $\xi(\alpha+\xi)^{-1}$. It is noteworthy that the steady state distribution of the server's status is insensitive to both the retrial rate θ and the Bernoulli routing parameter q.

Several other $M/G/1$-type retrial models can be analyzed as special cases of ours, and we present a few of these in the next section. We characterize the optimal Bernoulli routing parameter in section 5.

4 Some Special Cases

The last section revealed the insensitivity of the stability condition and the distribution of the server's status to the parameters θ and q. Interestingly, the stability condition and server status distribution for the present model are identical to those derived in [14]. However, equations (7)–(9) show that the retrial rate and Bernoulli routing parameter have a significant impact on the mean queue lengths. In this section, we demonstrate that our model can be used to analyze other $M/G/1$ retrial models. To this end, let S denote an arbitrary (uninterrupted) service time with c.d.f. B, and for $k \geq 1$, let

$$\beta_k \equiv \mathbb{E}(S^k) < \infty.$$

Denote by $\hat{\rho}$ the traffic intensity of the ordinary (non-retrial) $M/G/1$ queue with a perfectly reliable server so that $\hat{\rho} = \lambda\beta_1$, and suppose that $\hat{\rho} < 1$. Now, as $\xi \downarrow 0$ in (7)–(9), we obtain

$$\mathbb{E}(R) = \frac{\lambda^2 p}{1-\hat{\rho}}\left[\frac{\beta_1}{\theta} + \frac{\beta_2}{2(1-q\hat{\rho})}\right], \tag{10}$$

$$\mathbb{E}(Q) = \frac{\lambda^2 q\beta_2}{2(1-q\hat{\rho})}, \tag{11}$$

and

$$\mathbb{E}(N) = \hat{\rho} + \frac{\lambda^2}{1-\hat{\rho}}\left[\frac{\beta_1 p}{\theta} + \frac{\beta_2}{2}\right]. \tag{12}$$

Equations (10) and (11) are identical to equation (16) of Choi and Park [6] who studied a similar retrial model with a primary queue, retrial orbit, and a *reliable* server. Now, if we set $q = 0$ ($p = 1$) in (10)–(12), our model reduces to the standard $M/G/1$ retrial queue (i.e., one which does not possess a primary queue) with a reliable server, and the results are equivalent to those reported by Artalejo and Gómez-Corral [2] and Falin and Templeton [8]. Specifically,

$$\mathbb{E}(R) = \frac{\lambda^2}{1-\hat{\rho}}\left[\frac{\beta_1}{\theta} + \frac{\beta_2}{2}\right],$$

and

$$\mathbb{E}(N) = \hat{\rho} + \frac{\lambda^2}{1-\hat{\rho}}\left[\frac{\beta_1}{\theta} + \frac{\beta_2}{2}\right].$$

Note that $\mathbb{E}(Q) = 0$ here because all arriving customers who find the server busy are routed to the retrial orbit (i.e., there is no primary queue to accommodate waiting customers).

Next, consider our original model and allow $q \to 1$. With $q = 1$, we obtain the model analyzed by Sherman et al. [14] in which all arriving customers who find the server busy or failed join the primary queue. Allowing $q \to 1$ in (7)–(9) gives precisely the results obtained for $\mathbb{E}(R)$, $\mathbb{E}(Q)$, and $\mathbb{E}(N)$ in [14]. Moreover, if the uninterrupted service time distribution is exponential with parameter μ, we obtain the mean queue lengths of the model described in [13].

In the next section, we consider the problem of minimizing the total expected cost per unit time of holding customers in the primary queue and the orbit under a Bernoulli routing policy.

5 Optimal Bernoulli Routing Policy

We now consider the problem of determining an optimal Bernoulli routing policy for the retrial system with an unreliable server. The objective is to minimize the total expected holding costs per unit time. Under this criterion, we provide sufficient conditions to ensure the existence of a unique routing parameter, q^*. Denote the holding cost per customer per unit time in the primary queue by c_Q $(0 < c_Q < \infty)$, and let c_R $(0 < c_R < \infty)$ be the holding cost per customer per unit time in the retrial orbit. We consider only stable systems, i.e., those for which $\rho < 1$. Using (7) and (8), our optimization problem is of the form

$$\min \quad \vartheta(q) = c_R\mathbb{E}(R) + c_Q\mathbb{E}(Q) \tag{13a}$$

$$\text{s.t.} \quad q \in [0,1], \tag{13b}$$

where it is understood that $\mathbb{E}(R)$ and $\mathbb{E}(Q)$ depend explicitly on q. The optimal solution (when it exists) will be denoted by q^*. We elucidate the structure of $\vartheta(q)$ and characterize the optimal Bernoulli routing parameter in Proposition 4.

Proposition 4 *Suppose that $\rho < 1$ and $\xi > 0$. Then,*

(i) The cost function $\vartheta(q)$ is monotone increasing on $[0,1]$ if

$$\frac{c_Q}{c_R} - 1 > \frac{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)}{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}. \tag{14}$$

In this case, the optimal routing parameter is $q^* = 0$;

(ii) The cost function $\vartheta(q)$ is monotone decreasing on $[0,1]$ if

$$\frac{c_Q}{c_R} - 1 < \frac{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)(1-b^*(\xi)\rho)^2}{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}. \tag{15}$$

In this case, the optimal routing parameter is $q^* = 1$.

(iii) The cost function $\vartheta(q)$ is strictly convex on $[0,1]$ if

$$\begin{aligned}\frac{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)(1-b^*(\xi)\rho)^2}{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}} &\le \frac{c_Q}{c_R} - 1 \\ &\le \frac{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)}{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}.\end{aligned} \tag{16}$$

In this case, q^* uniquely solves $\vartheta'(q) = 0$ and is given by

$$q^* = \frac{1 - \sqrt{\left(\frac{c_Q}{c_R} - 1\right)\frac{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)}}}{b^*(\xi)\rho}. \tag{17}$$

Proof. Substituting $\mathbb{E}(R)$ and $\mathbb{E}(Q)$ from (7)–(8), respectively, and differentiating $\vartheta(q)$ respect to q, we obtain

$$\vartheta'(q) \equiv \frac{\mathrm{d}\vartheta(q)}{\mathrm{d}q} = (c_Q - c_R)\frac{\lambda\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2[\xi\hat{B}' - \lambda(1-b^*(\xi))]\right\}}{\alpha\xi^2 b^*(\xi)(\alpha+\xi)(1-qb^*(\xi)\rho)^2} - c_R\frac{\lambda(\xi+\alpha\rho)}{\theta\alpha(1-\rho)}. \tag{18}$$

The cost function $\vartheta(q)$ is (strictly) monotone increasing on $[0,1]$ if $\vartheta'(q) > 0$ for each $q \in [0,1]$, and in such case it attains its minimum at the boundary point 0. Equation (18) shows that $\vartheta'(q) > 0$ if

$$(c_Q - c_R)\frac{\lambda\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2[\xi\hat{B}' - \lambda(1-b^*(\xi))]\right\}}{\alpha\xi^2 b^*(\xi)(\alpha+\xi)(1-qb^*(\xi)\rho)^2} > c_R\frac{\lambda(\xi+\alpha\rho)}{\theta\alpha(1-\rho)},$$

or equivalently, if

$$\frac{c_Q}{c_R} - 1 > \frac{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)}{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}.$$

Similarly, $\vartheta(q)$ is (strictly) monotone decreasing on $[0,1]$ if $\vartheta'(q) < 0$ for each $q \in [0,1]$, and in this case $\vartheta(q)$ attains its minimum at 1. By (18), this condition is met if

$$\frac{c_Q}{c_R} - 1 < \frac{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)(1-b^*(\xi)\rho)^2}{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}.$$

Finally, we prove that there is a region for which the strict convexity of $\vartheta(q)$ is ensured. Differentiating (18) with respect to q yields

$$\vartheta''(q) \equiv \frac{\mathrm{d}^2\vartheta(q)}{\mathrm{d}q^2} = (c_Q - c_R)\frac{2\lambda\rho\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2[\xi\hat{B}' - \lambda(1-b^*(\xi))]\right\}}{\alpha\xi^2(\alpha+\xi)(1-qb^*(\xi)\rho)^3}. \tag{19}$$

The denominator of (19) is strictly positive for any $q \in [0,1]$ and $\rho < 1$ since $0 < b^*(\xi) < 1$ for any $\xi > 0$. We can conclude that the numerator is also strictly positive since it was shown in the proof of Lemma 2 of [14] that $\lambda(1-b^*(\xi)) - \xi\hat{B}' \geq 0$ for $\xi \geq 0$. By rearranging the terms of (16), it is seen that c_Q must exceed c_R in this region. Therefore, we conclude that when condition (16) is met, $\vartheta''(q) > 0$ for each $q \in [0,1]$, and hence, $\vartheta(q)$ is strictly convex and has a unique stationary point q^* satisfying $\vartheta'(q^*) = 0$. By (18), this point is given by

$$q^* = \frac{1 - \sqrt{\left(\frac{c_Q}{c_R} - 1\right)\frac{\theta(1-\rho)\left\{\xi^3 b^*(\xi) - (\alpha+\xi)^2\left[\xi\hat{B}' - \lambda(1-b^*(\xi))\right]\right\}}{\xi^2 b^*(\xi)(\alpha+\xi)(\xi+\alpha\rho)}}}{b^*(\xi)\rho}. \tag{20}$$

■

The advantage of Proposition 4 is that it allows us to determine the form of the optimal solution by simply checking the value $c_Q/c_R - 1$. Section 6 illustrates and highlights the main results through a few numerical examples.

6 Numerical Examples

In this section, we illustrate the behavior of the cost function $\vartheta(q)$ in the three regimes identified by Proposition 4. Additionally, we provide a comparison of the derived mean performance parameters (namely the mean queue lengths) with values obtained via a discrete-event simulation model. For both illustrations we consider two absolutely continuous service time distributions whose Laplace transforms (LTs) are well defined, namely the exponential and uniform distributions. To compute the mean queue lengths and the cost function, we need expressions for $b^*(\xi)$, the LT of the service time distribution evaluated at ξ, and $\hat{B}'$ which is given by

$$\hat{B}' = \lambda \int_0^\infty x \exp(-\xi x) b(x) \mathrm{d}x.$$

Let S denote an arbitrary (uninterrupted) service time. When S is distributed exponentially with rate μ, it is easy to show that

$$b^*(\xi) = \frac{\mu}{\mu + \xi},$$

and

$$\hat{B}' = \frac{\lambda\mu}{(\mu + \xi)^2}.$$

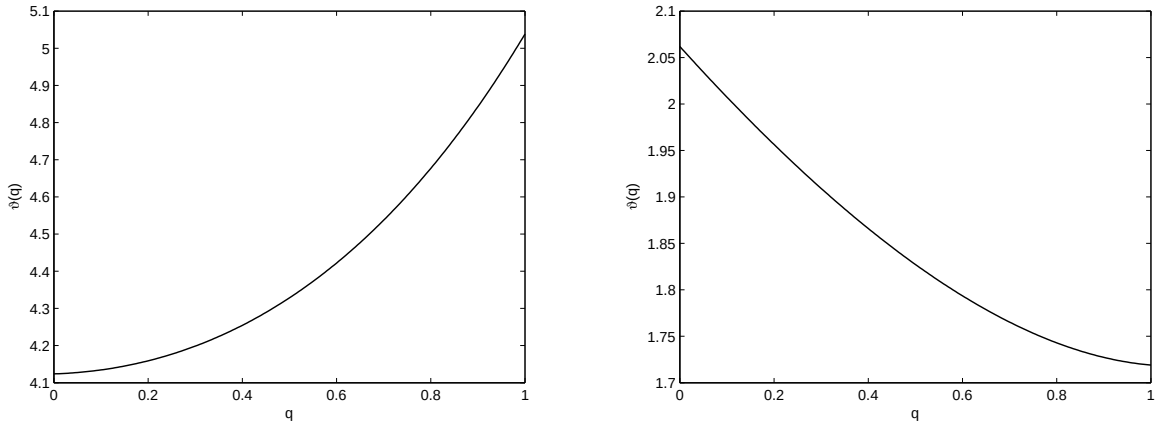
When S is distributed uniformly on the interval $(0, y)$, $0 < y < \infty$, we obtain

$$b^*(\xi) = \frac{1}{y\xi} \left[1 - \exp(-y\xi)\right],$$

and

$$\hat{B}' = \frac{\lambda}{y\xi^2} \left[(1 - \exp(-y\xi)(1 + y\xi)\right].$$

Before comparing mean queue lengths with simulated results, we first illustrate the cost function, $\vartheta(q)$. For both distributions we use the following parameters: $\lambda = 2.0$, $\mu = 10.0$, $\xi = 1.0$, $\alpha = 2.0$, and $\theta = 2.0$. In the case of exponential service times, the traffic intensity is $\rho \approx 0.3100$. Figure 1 depicts two cases when the cost function $\vartheta(q)$ is monotone. In Figure 1(a), the cost parameters are $c_Q = 8.0$ and $c_R = 2.0$ so that $c_Q/c_R - 1 = 3.0$, and the cost function is monotone increasing on $[0, 1]$ as dictated by part (i) of Proposition 4. For Figure 1(b), we use $c_Q = 2.5$ and $c_R = 1$ so that $c_Q/c_R - 1 = 1.50$, and the cost function is monotone decreasing on $[0, 1]$ in accordance with part (ii) of Proposition 4.



(a) Exponential case (i): $\vartheta(q)$ is monotone increasing.

(b) Exponential case (ii): $\vartheta(q)$ is monotone decreasing.

Figure 1: Sample monotone cost functions when service time is exponential.

In case (i), the optimal Bernoulli routing parameter and corresponding minimum cost are given by $q^* = 0.0$ and $\vartheta(q^*) \approx 4.12381$, respectively. Using this set of parameter values and holding cost coefficients, it is optimal to divert all primary arrivals to the orbit when the server is busy or failed. For case (ii), the optimal Bernoulli routing parameter and corresponding minimum cost are $q^* = 1.0$ and $\vartheta(q^*) \approx 1.71905$, respectively. In this case, it is optimal to admit all new arrivals to the primary queue when the server is busy or failed, despite the fact that the holding cost in the primary queue is more than double that of the retrial queue. This case illustrates the fact that the primary queue can remain stable even as the retrial queue continues to grow and shows that the relative magnitudes of the cost coefficients are not the only determinants of the optimal Bernoulli routing policy.

Figure 2 depicts the cost function when $c_Q = 3$, $c_R = 1$, and $c_Q/c_R - 1 = 2$. The cost function is strictly convex on $[0, 1]$ in accordance with part (iii) of Proposition 4. The optimal Bernoulli routing parameter, computed by equation (17), and the corresponding minimum cost are given by

$$q^* \approx 0.6450 \quad \text{and} \quad \vartheta(q^*) \approx 1.9319,$$

respectively. In this case, we see that there is a tradeoff between the cost coefficients and the growth of the queue lengths. In particular, it is optimal for the controller to divert roughly 35% of the arrival stream directly to the orbit while nearly 65% are admitted to the primary queue.

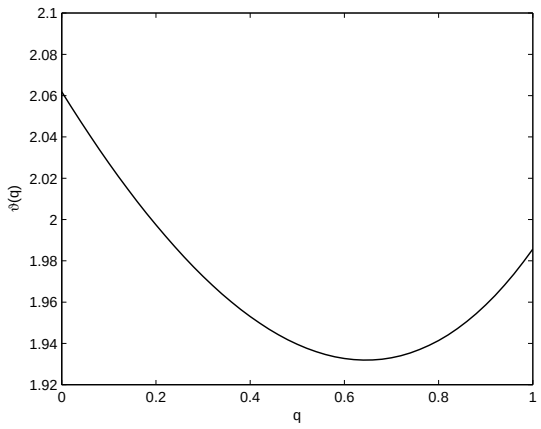
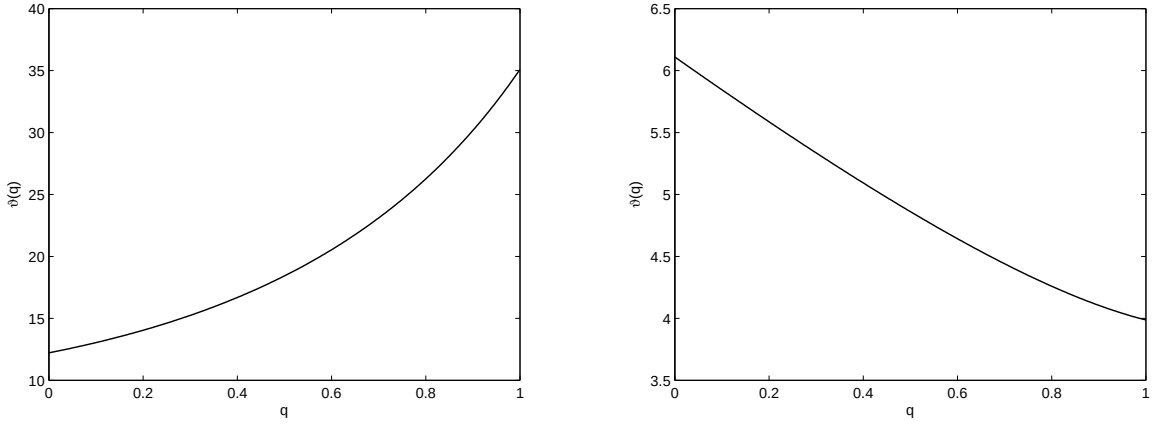


Figure 2: Exponential case (iii): $\vartheta(q)$ is strictly convex.

Next, we consider the case when the service time S is uniformly distributed on $(0, y)$ where we assume $y = 4/\mu$, i.e., the mean service time is twice that of the exponential case with all other parameter values unchanged. For this case, the traffic intensity is $\rho \approx 0.63989$. Similar graphs can be drawn for this case; however, the cost coefficients need to be altered to conform to the conditions of Proposition 4.

Figure 3 depicts two cases when the cost function $\vartheta(q)$ is monotone. In Figure 3(a), the cost parameters are $c_Q = 30$ and $c_R = 2$ so that $c_Q/c_R - 1 = 14$, and the cost function is monotone increasing on $[0, 1]$ as dictated by part (i) of Proposition 4. For Figure 3(b), we use $c_Q = 2$ and $c_R = 1$ so that $c_Q/c_R - 1 = 1$, and the cost function is monotone decreasing on $[0, 1]$ in accordance with part (ii) of Proposition 4. In case (i), we obtain $q^* = 0$ and $\vartheta(q^*) \approx 12.21908$, while in case (ii), the optimal routing parameter is $q^* = 1$ with a corresponding minimum cost of $\vartheta(q^*) \approx 3.98708$.



(a) Uniform case (i): $\vartheta(q)$ is monotone increasing.

(b) Uniform case (ii): $\vartheta(q)$ is monotone decreasing.

Figure 3: Sample monotone cost functions when service time is uniform.

Finally, in Figure 4, we use $c_Q = 3$ and $c_R = 1$ so that $c_Q/c_R - 1 = 2$, and the cost function

is strictly convex on $[0, 1]$ in accordance with part (iii) of Proposition 4. The optimal Bernoulli routing parameter, and corresponding minimum cost, are given by $q^* \approx 0.8400$ and $\vartheta(q^*) \approx 4.93731$, respectively.

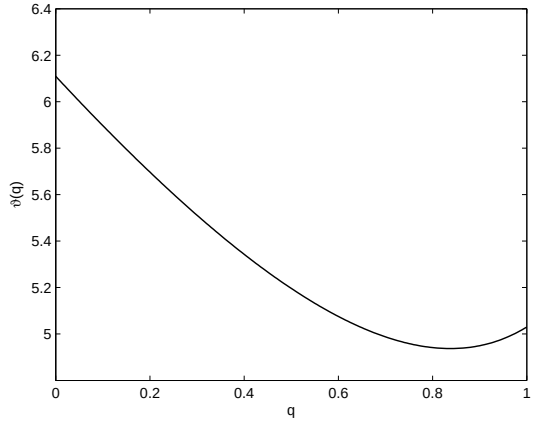


Figure 4: Uniform case (iii): $\vartheta(q)$ is strictly convex.

As expected, it is seen that the optimal Bernoulli routing parameter that minimizes the total expected holding costs per unit time is on the interior of the feasible region.

Next, we examine the mean queue lengths as a function of the overall traffic intensity, ρ. In the experiments that follow, we assumed $\mu = 3$, $\alpha = 4$, $\xi = 0.1$, $\theta = 5$, $q = 0.5$, and we varied λ to obtain increasing values of ρ. Table 1 shows a comparison of the computed mean queue lengths as compared to the same values obtained via a discrete-event simulation model when the service time is exponential with parameter $\mu = 3$.

Table 1: Mean queue lengths with exponential service times.

	$\mathbb{E}(R)$		$\mathbb{E}(Q)$		$\mathbb{E}(N)$	
ρ	Analytical	Simulated	Analytical	Simulated	Analytical	Simulated
0.1	0.015922	0.015890	0.005898	0.005920	0.119381	0.119470
0.3	0.140043	0.139810	0.052832	0.052670	0.485558	0.485170
0.5	0.542132	0.541680	0.161568	0.161440	1.191505	1.190800
0.7	1.874600	1.868800	0.359237	0.358610	2.916764	2.910300
0.9	10.207577	10.173000	0.691578	0.690730	11.777204	11.741000

It is interesting to note in Table 1 that there is a dramatic increase in $\mathbb{E}(R)$ (and $\mathbb{E}(N)$) when the traffic intensity increases from 0.7 to 0.9. By contrast, the increase in the mean primary queue size, $\mathbb{E}(Q)$, is more moderate.

For the case of uniformly-distributed service times, the parameter values are the same; however, the service time is assumed to be $U(0, 2/\mu)$ where μ is the rate parameter of the exponential case.

Note that the mean service time is identical to that of the exponential case. Table 2 summarizes the steady state mean queue lengths and mean number of customers in the system. In this exper-

Table 2: Mean queue lengths with uniform service times.

	$\mathbb{E}(R)$		$\mathbb{E}(Q)$		$\mathbb{E}(N)$	
ρ	Analytical	Simulated	Analytical	Simulated	Analytical	Simulated
0.1	0.013951	0.013950	0.004270	0.004270	0.115781	0.115750
0.3	0.114872	0.114950	0.036583	0.036580	0.444137	0.444220
0.5	0.432099	0.432210	0.110704	0.110660	1.030608	1.030600
0.7	1.463517	1.463800	0.244981	0.244980	2.391425	2.391500
0.9	7.824286	7.830100	0.470328	0.470290	9.172663	9.178500

iment, we also note that, as the overall traffic intensity increases, there is a much more profound effect on $\mathbb{E}(R)$ and $E(N)$ than on the mean primary queue length whose stability depends on ρ_1. Specifically, when $\rho = 0.90$, $\rho_1 \approx 0.4353$. Therefore, the increase in the overall traffic intensity from 0.7 to 0.9 results in only a moderate increase in the primary queue length.

Acknowledgements

The authors are grateful to two anonymous referees and Professor Sheldon Ross for helpful comments and suggestions that have improved an earlier version of this paper. This work was sponsored, in part, by a grant from the Air Force Office of Scientific Research (FA9550-08-1-0004).

References

[1] Aissani, A. and Artalejo, J.R. (1998). On the single server retrial queue subject to breakdowns. *Queueing Systems,* 30(3-4): 309-321.

[2] Artalejo, J.R. and Gómez-Corral, A. (2008). *Retrial Queueing Systems: A Computational Approach*. Springer-Verlag, Berlin, Germany.

[3] Atencia, I. and Moreno, P. (2005). A single-server retrial queue with general retrial times and Bernoulli schedule. *Applied Mathematics and Computation,* 162(2): 855-880.

[4] Casetti, C., Cigno, R.L. and Mellia, M. (2000). Load-balancing solutions for static routing schemes in ATM networks. *Computer Networks*, 34: 169-180.

[5] Chang, C-S, Chao, X. and Pinedo, M. (1990). A note on queues with Bernoulli routing. In *Proceedings of the 29th Conference on Decision and Control*, Honolulu, Hawaii, December 1990, pp. 897-902.

[6] Choi, B.D. and Park, K.K. (1990). The $M/G/1$ retrial queue with Bernoulli schedule. *Queueing Systems*, 7(2): 219-228.

[7] Combé, M.B. and Boxma, O.J. (1994). Optimization of static traffic allocation policies. *Theoretical Computer Science*, 125(1): 17-43.

[8] Falin, G.I. and Templeton, J.G.C. (1997). *Retrial Queues*. Chapman and Hall, London.

[9] Falin, G.I. (2008). The $M/M/1$ retrial queue with retrials due to failures. *Queueing Systems*, 58: 155-160.

[10] Koole, G. (1996).On the pathwise optimal Bernoulli routing policy for homogeneous parallel servers. *Mathematics of Operations Research*, 21(2): 469-476.

[11] Liang, H.M. and Kulkarni, V.G. (1999). Optimal routing control in retrial queues. In *Applied Probability and Stochastic Processes: International Series in Operations Research and Management Science*, Volume 19, (J.G. Shanthikumar and Ushio Sumita, eds.), pp. 203-218.

[12] Servi, L. and Humair, S. (1999). Optimizing Bernoulli routing policies for balancing loads on call centers and minimizing transmission costs. *Journal of Optimization Theory and Applications*, 100(3): 623-659.

[13] Sherman, N.P. and Kharoufeh, J.P. (2006). An $M/M/1$ retrial queue with unreliable server. *Operations Research Letters*, 34(6): 697-705.

[14] Sherman, N.P., Kharoufeh, J.P. and Abramson, M. (2009). An $M/G/1$ retrial queue with unreliable server for streaming multimedia applications. *Probability in the Engineering and Informational Sciences*, 23(2): 281-304.